

## CPE232 Data Models

### Introduction

โปรเจกต์นี้เป็นการนำข้อมูล จากคลังข้อมูลสาธารณะของ GHO (Global Health Observatory) ภายใต้องค์การอนามัยโลก (WHO) เพื่อติดตามสถานะด้านสุขภาพและปัจจัยที่เกี่ยวข้องอื่นๆ สำหรับทุกประเทศ ข้อมูลเหล่านี้ถูกเผยแพร่ให้สาธารณะเพื่อการวิเคราะห์ข้อมูลทางสุขภาพ ข้อมูลเกี่ยวกับอายุขัยชีวิตและปัจจัยสุขภาพสำหรับ 193 ประเทศได้รับการเก็บรวบรวมจากเว็บไซต์เดียวกันของคลังข้อมูล WHO และข้อมูลเศรษฐกิจที่เกี่ยวข้องถูกเก็บรวบรวมจากเว็บไซต์ของสหประชาชาติ จากหมวดหมู่ที่เกี่ยวข้องกับปัจจัยสุขภาพทั้งหมด ในช่วง 15 ปี ตั้งแต่ปี 2000-2015 จำนวน 193 ประเทศเพื่อวิเคราะห์ และทำนายอายุจากข้อมูลที่มี

Data explanation each column

Column Name	Description
Country	ชื่อประเทศ
Year	ปีที่เก็บข้อมูล
Status	เป็นประเทศที่พัฒนาแล้ว หรือ กำลังพัฒนา
Life expectancy	การคาดประมาณจำนวนอายุโดยเฉลี่ยของการมีชีวิตอยู่ของประชากร
Adult Mortality	อัตราการเสียชีวิตของคนอายุ 15-60 ต่อ 1000 คน
Infant Deaths	จำนวนทารกแรกเกิดที่เสียชีวิต ต่อ 1000 คน
Alcohol	จำนวนการบริโภคแอลกอฮอล์ เป็นลิตรต่อคน (อายุ 15 ขึ้นไป)
Percentage Expenditure	ค่าใช้จ่ายด้านสุขภาพคิดเป็นเปอร์เซ็นต์ของผลิตภัณฑ์รวมในประเทศต่อหัว (%)
Hepatitis B	ความครอบคลุมการฉีดวัคซีนป้องกันไวรัสตับอักเสบบี (HepB) ในเด็กอายุ 1 ปี (%)
Measles	จำนวนรายงานผู้ป่วยโรคหัดต่อประชากร 1,000 คน
Bmi	ดัชนีมวลกายเฉลี่ยของประชากรทั้งหมด
Under-Five Deaths	อัตราการเสียชีวิตอายุที่ต่ำกว่า 5 ปีต่อประชากร 1,000 คน
Polio	ความครอบคลุมการฉีดวัคซีนป้องกันโรคโปลิโอ ในเด็กอายุ 1 ปี (%)
Total Expenditure	รายจ่ายด้านสุขภาพของรัฐบาลคิดเป็นเปอร์เซ็นต์ของรายจ่ายภาครัฐทั้งหมด (%)
Diphtheria	ความครอบคลุมการฉีดวัคซีนป้องกันโรคคอตีบ บาดทะยัก และไอกรน ในเด็กอายุ 1 ปี (%)
Hiv/Aids	อัตราการเสียชีวิตต่อการเกิดเอชไอวี/เอดส์ที่ต่ำกว่า 5 ปี ต่อประชากร 1,000 คน
Gdp	ผลิตภัณฑ์มวลรวมภายในประเทศต่อคน (หน่วยเป็น USD)
Population	จำนวนประชากร
Thinness 10-19 Years	ภาวะผอมของเด็กอายุ 10 - 19 ปี

Thisness 5-9 Years	ภาวะผอมของเด็กอายุ 5 - 9 ปี
Income Composition Of Resources	ดัชนีการพัฒนามนุษย์ในแง่ของรายได้ของทรัพยากร (ดัชนีตั้งแต่ 0 ถึง 1)
Schooling	จำนวนปีการศึกษา(ปี)

## Data preparation process and results

```
import pandas as pd
import numpy as np
```

ขั้นแรกเราจะทำการ import library ที่เราต้องการที่จะใช้มาก่อน

```
data = pd.read_csv(r"C:\kmutt\2.2\CPE232\finalpj\Life Expectancy Data.csv")
pd.set_option('display.max_columns', None)
data.head()
```

ต่อไปจะอ่านข้อมูลจากไฟล์ CSV เก็บข้อมูลไว้ใน data ตั้งค่าให้แสดงทุกคอลัมน์ของข้อมูลและแสดงข้อมูลหน้าต่อหน้าของตาราง

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	19.1	83	6.0	8.16	65.0	0.1	504.259210	33736494.0	17.2	17.3	0.479	10.1
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	18.6	86	58.0	8.18	62.0	0.1	612.696514	327582.0	17.5	17.5	0.476	10.0
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	18.1	89	62.0	8.13	64.0	0.1	631.744976	31731688.0	17.7	17.7	0.470	9.9
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	17.6	93	67.0	8.52	67.0	0.1	669.959000	3696958.0	17.9	18.0	0.463	9.8
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	17.2	97	68.0	7.87	68.0	0.1	63.537231	2978599.0	18.2	18.2	0.454	9.5

อันนี้เป็น output ของ data ที่เรานั้นได้มา

```
data.info()
```

ต่อไปจะดูข้อมูลทางเทคนิคของ dataframe เช่น จำนวนแถวและคอลัมน์ทั้งหมด ชนิดของข้อมูลในแต่ละคอลัมน์

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Country                                   2938 non-null   object
1   Year                                       2938 non-null   int64
2   Status                                    2938 non-null   object
3   Life expectancy                          2928 non-null   float64
4   Adult Mortality                          2928 non-null   float64
5   infant deaths                            2938 non-null   int64
6   Alcohol                                   2744 non-null   float64
7   percentage expenditure                   2938 non-null   float64
8   Hepatitis B                             2385 non-null   float64
9   Measles                                  2938 non-null   int64
10  BMI                                       2904 non-null   float64
11  under-five deaths                       2938 non-null   int64
12  Polio                                    2919 non-null   float64
13  Total expenditure                       2712 non-null   float64
14  Diphtheria                             2919 non-null   float64
15  HIV/AIDS                               2938 non-null   float64
16  GDP                                      2490 non-null   float64
17  Population                              2286 non-null   float64
18  thinness 1-19 years                     2904 non-null   float64
19  thinness 5-9 years                      2904 non-null   float64
20  Income composition of resources         2771 non-null   float64
21  Schooling                               2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB

```

ข้อมูลที่ได้จาก data.info()

```
data.isnull().any()
```

ต่อไปเราจะใช้ data.isnull().any() เพื่อดูว่าข้อมูลของเรานั้น มี column ไหนบ้างที่มี NULL โดยที่ output ของเราจะได้ออกมาในรูปแบบของ boolean

```
Country      False
Year         False
Status       False
Life expectancy  True
Adult Mortality  True
infant deaths  False
Alcohol      True
percentage expenditure  False
Hepatitis B   True
Measles      False
BMI          True
under-five deaths  False
Polio        True
Total expenditure  True
Diphtheria   True
HIV/AIDS    False
GDP          True
Population   True
thinness 1-19 years  True
thinness 5-9 years  True
Income composition of resources  True
Schooling    True
dtype: bool
```

ข้อมูลที่ได้จาก data.isnull().any()

```
data[data['Life expectancy '].isnull()]
```

หลังจากนั้นเราจะมาดูตารางในแถวที่มีค่า Life expectancy เป็น NULL ว่ามันเกิดจากอะไร

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
624	Cook Islands	2013	Developing	NaN	NaN	0	0.01	0.000000	98.0	0	82.8	0	98.0	3.58	98.0	0.1	NaN	NaN	0.1	0.1	NaN	NaN
769	Dominica	2013	Developing	NaN	NaN	0	0.01	11.419555	96.0	0	58.4	0	96.0	5.58	96.0	0.1	722.756650	NaN	2.7	2.6	0.721	12.7
1650	Marshall Islands	2013	Developing	NaN	NaN	0	0.01	871.878317	8.0	0	81.6	0	79.0	17.24	79.0	0.1	3617.752354	NaN	0.1	0.1	NaN	0.0
1715	Monaco	2013	Developing	NaN	NaN	0	0.01	0.000000	99.0	0	NaN	0	99.0	4.30	99.0	0.1	NaN	NaN	NaN	NaN	NaN	NaN
1812	Nauru	2013	Developing	NaN	NaN	0	0.01	15.606596	87.0	0	87.3	0	87.0	4.65	87.0	0.1	136.183210	NaN	0.1	0.1	NaN	9.6
1909	Niue	2013	Developing	NaN	NaN	0	0.01	0.000000	99.0	0	77.3	0	99.0	7.20	99.0	0.1	NaN	NaN	0.1	0.1	NaN	NaN
1958	Palau	2013	Developing	NaN	NaN	0	NaN	344.690631	99.0	0	83.3	0	99.0	9.27	99.0	0.1	1932.122370	292.0	0.1	0.1	0.779	14.2
2167	Saint Kitts and Nevis	2013	Developing	NaN	NaN	0	8.54	0.000000	97.0	0	5.2	0	96.0	6.14	96.0	0.1	NaN	NaN	3.7	3.6	0.749	13.4
2216	San Marino	2013	Developing	NaN	NaN	0	0.01	0.000000	69.0	0	NaN	0	69.0	6.50	69.0	0.1	NaN	NaN	NaN	NaN	NaN	15.1
2713	Tuvalu	2013	Developing	NaN	NaN	0	0.01	78.281203	9.0	0	79.3	0	9.0	16.61	9.0	0.1	3542.136980	1619.0	0.2	0.1	NaN	0.0

เมื่อเราได้ข้อมูลออกมา เราก็จะทำการวิเคราะห์ข้อมูลที่เราได้ออกมาว่าเราควรจะทำอย่างไร โดยที่แถวที่ Life expectancy เป็น NULL นั้น มันเป็น NaN ซึ่งค่า Life expectancy นั้น จะเป็นค่าที่เราเอามาใช้เป็นค่าที่เราจะนำมาเรียนรู้เพื่อที่จะทำนาย เราจึงตัดสินใจที่จะลบแถวที่ Life expectancy นั้นเป็น NULL ทิ้งไป

```
new=data[data['Life expectancy '].isnull()]
data.drop(new.index,inplace=True)
```

โค้ดนี้จะเป็นการลบแถวที่ Life expectancy นั้นเป็น NULL ทิ้งไป โดยที่จะสร้าง dataframe ใหม่มาอันนึง

โดยเลือกแถวที่มีคอลัมน์ Life expectancy เป็น NULL จากนั้นจะ drop แถวใน data ที่มี index เดียวกันกับ new ออกไปโดยไม่ต้องกำหนด DataFrame ใหม่

```
data.isnull().any()
```

ต่อไปก็ดู column ที่เป็น NULL อีกรอบ

Country	False
Year	False
Status	False
Life expectancy	False
Adult Mortality	False
infant deaths	False
Alcohol	True
percentage expenditure	False
Hepatitis B	True
Measles	False
BMI	True
under-five deaths	False
Polio	True
Total expenditure	True
Diphtheria	True
HIV/AIDS	False
GDP	True
Population	True
thinness 1-19 years	True
thinness 5-9 years	True
Income composition of resources	True
Schooling	True
dtype: bool	

ข้อมูลที่ได้จาก data.isnull().any()

```
data[data['Alcohol'].isnull()]
```

ต่อไปจะดูตารางในแถวที่มีค่า Alcohol เป็น NULL ว่ามันเกิดจากอะไร

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
32	Algeria	2015	Developing	75.6	19.0	21	NaN	0.0	95.0	63	59.5	24	95.0	NaN	95.0	0.1	4132.762920	39871528.0	6.0	5.8	0.743	14.4
48	Angola	2015	Developing	52.4	335.0	66	NaN	0.0	64.0	118	23.3	98	7.0	NaN	64.0	1.9	3695.793748	2785935.0	8.3	8.2	0.531	11.4
64	Antigua and Barbuda	2015	Developing	76.4	13.0	0	NaN	0.0	99.0	0	47.7	0	86.0	NaN	99.0	0.2	13566.954100	NaN	3.3	3.3	0.784	13.9
80	Argentina	2015	Developing	76.3	116.0	8	NaN	0.0	94.0	0	62.8	9	93.0	NaN	94.0	0.1	13467.123600	43417765.0	1.0	0.9	0.826	17.3
96	Armenia	2015	Developing	74.8	118.0	1	NaN	0.0	94.0	33	54.9	1	96.0	NaN	94.0	0.1	369.654776	291695.0	2.1	2.2	0.741	12.7
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2858	Venezuela (Bolivarian Republic of)	2015	Developing	74.1	157.0	9	NaN	0.0	87.0	0	62.1	10	87.0	NaN	87.0	0.1	NaN	NaN	1.6	1.5	0.769	14.3
2874	Viet Nam	2015	Developing	76.0	127.0	28	NaN	0.0	97.0	256	17.5	35	97.0	NaN	97.0	0.1	NaN	NaN	14.2	14.5	0.678	12.6
2890	Yemen	2015	Developing	65.7	224.0	37	NaN	0.0	69.0	468	41.3	47	63.0	NaN	69.0	0.1	NaN	NaN	13.6	13.4	0.499	9.0
2906	Zambia	2015	Developing	61.8	33.0	27	NaN	0.0	9.0	9	23.4	40	9.0	NaN	9.0	4.1	1313.889646	161587.0	6.3	6.1	0.576	12.5
2922	Zimbabwe	2015	Developing	67.0	336.0	22	NaN	0.0	87.0	0	31.8	32	88.0	NaN	87.0	6.2	118.693830	15777451.0	5.6	5.5	0.507	10.3

193 rows x 22 columns

ต่อไปเป็นการวิเคราะห์ข้อมูลแถวที่มีค่า Alcohol เป็น NULL โดยที่เราได้ว่า column Alcohol นั้นใช้อธิบายหน่วยการบริโภคแอลกอฮอล์ต่อคน ซึ่งในตารางนี้ไม่มีค่าที่เป็น 0.0 ซึ่งเป็นไปไม่ได้ที่จะไม่มีคนที่ไม่บริโภคแอลกอฮอล์ ดังนั้นเราจึงเลือกวิธีการเปลี่ยนเป็น 0 ในช่องที่เป็น NULL

```
data['Alcohol'].fillna('0.0', inplace=True)
```

โค้ดนี้เป็นการเติมค่า NULL ในคอลัมน์ Alcohol ด้วยค่า 0.0

```
data[data['Hepatitis B'].isnull()]
```

ต่อไปจะดูตารางในแถวที่มีค่า Hepatitis B เป็น NULL ว่ามันเกิดจากอะไร

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
44	Algeria	2003	Developing	71.7	146.0	20	0.34	25.018523	NaN	15374	47.0	23	87.0	3.60	87.0	0.1	294.335560	3243514.0	6.3	6.1	0.663	11.5
45	Algeria	2002	Developing	71.6	145.0	20	0.36	148.511984	NaN	5862	46.1	23	86.0	3.73	86.0	0.1	1774.336730	3199546.0	6.3	6.2	0.653	11.1
46	Algeria	2001	Developing	71.4	145.0	20	0.23	147.986071	NaN	2686	45.3	24	89.0	3.84	89.0	0.1	1732.857979	31592153.0	6.4	6.3	0.644	10.9
47	Algeria	2000	Developing	71.3	145.0	21	0.25	154.455944	NaN	0	44.4	25	86.0	3.49	86.0	0.1	1757.177970	3118366.0	6.5	6.4	0.636	10.7
57	Angola	2006	Developing	47.7	381.0	90	5.84	25.086888	NaN	765	18.2	143	36.0	4.54	34.0	2.5	262.415149	2262399.0	9.8	9.7	0.439	7.2
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2917	Zambia	2004	Developing	47.9	578.0	36	2.46	8.369852	NaN	35	18.0	59	84.0	7.33	83.0	17.6	53.277222	11731746.0	7.2	7.1	0.456	10.5
2918	Zambia	2003	Developing	46.4	64.0	39	2.33	65.789974	NaN	881	17.6	62	85.0	8.18	83.0	18.2	429.168343	11421984.0	7.3	7.2	0.443	10.2
2919	Zambia	2002	Developing	45.5	69.0	41	2.44	54.043480	NaN	25036	17.3	66	85.0	6.93	84.0	18.4	377.135244	111249.0	7.4	7.3	0.433	10.0
2920	Zambia	2001	Developing	44.6	611.0	43	2.61	46.830275	NaN	16997	17.1	70	86.0	6.56	85.0	18.6	378.273624	1024125.0	7.4	7.4	0.424	9.8
2921	Zambia	2000	Developing	43.8	614.0	44	2.62	45.616880	NaN	30930	16.8	72	85.0	7.16	85.0	18.7	341.955625	1531221.0	7.5	7.5	0.418	9.6

553 rows x 22 columns

ต่อไปเป็นการวิเคราะห์ข้อมูลแถวที่มีค่า Hepatitis B เป็น NULL โดยที่เราได้ว่า column Hepatitis B นั้นใช้อธิบายหน่วยการการฉีดวัคซีนป้องกันไวรัสตับอักเสบบีในเด็กอายุ 1 ปีคิดเป็นเปอร์เซ็นต์ ซึ่งในตารางนี้ไม่มีค่าที่เป็น 0.0 ซึ่งมันสามารถมีการฉีดวัคซีนป้องกันไวรัสตับอักเสบบีเป็น 0 เปอร์เซนต์ได้ ดังนั้นเราจึงเลือกวิธีการเปลี่ยนเป็น 0 ในช่องที่เป็น NULL

```
data['Hepatitis B'].fillna('0.0', inplace=True)
```

โค้ดนี้เป็นการเติมค่า NULL ในคอลัมน์ Hepatitis B ด้วยค่า 0.0

```
data.isnull().any()
```

ต่อไปก็ดู column ที่เป็น NULL อีกรอบ



```

Country                False
Year                   False
Status                 False
Life expectancy        False
Adult Mortality        False
infant deaths          False
Alcohol                False
percentage expenditure False
Hepatitis B            False
Measles                False
    BMI                 True
under-five deaths      False
Polio                  True
Total expenditure      True
Diphtheria             True
    HIV/AIDS            False
GDP                    True
Population              True
    thinness 1-19 years  True
    thinness 5-9 years  True
Income composition of resources True
Schooling               True
dtype: bool

```

ข้อมูลที่ได้จาก data.isnull().any()

```
data[data[' BMI '].isnull()]
```

ต่อไปจะดูตารางในแถวที่มีค่า BMI เป็น NULL ว่ามันเกิดจากอะไร

	Country	Year	Status	Life expectancy	Adult Mortality	Infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
2409	South Sudan	2015	Developing	57.3	332.0	26	0.0	0.000000	31.0	878	NaN	39	41.0	NaN	31.0	3.4	758.725782	11882136.0	NaN	NaN	0.421	4.9
2410	South Sudan	2014	Developing	56.6	343.0	26	0.0	46.074469	0.0	441	NaN	39	44.0	2.74	39.0	3.5	1151.861715	1153971.0	NaN	NaN	0.421	4.9
2411	South Sudan	2013	Developing	56.4	345.0	26	0.0	47.444530	0.0	525	NaN	40	5.0	2.62	45.0	3.6	1186.113250	1117749.0	NaN	NaN	0.417	4.9
2412	South Sudan	2012	Developing	56.0	347.0	26	0.0	38.338232	0.0	1952	NaN	40	64.0	2.77	59.0	3.8	958.455810	1818258.0	NaN	NaN	0.419	4.9
2413	South Sudan	2011	Developing	55.4	355.0	27	0.0	0.000000	0.0	1256	NaN	41	66.0	NaN	61.0	3.9	176.971300	1448857.0	NaN	NaN	0.429	4.9
2414	South Sudan	2010	Developing	55.0	359.0	27	0.0	0.000000	0.0	0	NaN	41	NaN	NaN	NaN	4.0	1562.239346	167192.0	NaN	NaN	0.000	0.0
2415	South Sudan	2009	Developing	54.3	369.0	27	0.0	0.000000	0.0	0	NaN	42	NaN	NaN	NaN	4.2	1264.789980	967667.0	NaN	NaN	0.000	0.0
2416	South Sudan	2008	Developing	53.6	377.0	27	0.0	0.000000	0.0	0	NaN	42	NaN	NaN	NaN	4.2	1678.711862	9263136.0	NaN	NaN	0.000	0.0
2417	South Sudan	2007	Developing	53.1	381.0	27	0.0	0.000000	0.0	0	NaN	43	NaN	NaN	NaN	4.2	NaN	88568.0	NaN	NaN	0.000	0.0
2418	South Sudan	2006	Developing	52.5	383.0	28	0.0	0.000000	0.0	0	NaN	43	NaN	NaN	NaN	4.1	NaN	8468152.0	NaN	NaN	0.000	0.0
2419	South Sudan	2005	Developing	51.9	383.0	28	0.0	0.000000	0.0	0	NaN	44	NaN	NaN	NaN	3.9	NaN	818877.0	NaN	NaN	0.000	0.0
2420	South Sudan	2004	Developing	51.4	383.0	29	0.0	0.000000	0.0	0	NaN	45	NaN	NaN	NaN	3.8	NaN	7787655.0	NaN	NaN	0.000	0.0
2421	South Sudan	2003	Developing	58.0	383.0	29	0.0	0.000000	0.0	0	NaN	46	NaN	NaN	NaN	3.5	NaN	751642.0	NaN	NaN	0.000	0.0
2422	South Sudan	2002	Developing	52.0	382.0	30	0.0	0.000000	0.0	0	NaN	48	NaN	NaN	NaN	3.3	NaN	7237276.0	NaN	NaN	0.000	0.0
2423	South Sudan	2001	Developing	49.6	381.0	30	0.0	0.000000	0.0	0	NaN	49	NaN	NaN	NaN	3.0	NaN	6974442.0	NaN	NaN	0.000	0.0
2424	South Sudan	2000	Developing	48.9	38.0	31	0.0	0.000000	0.0	0	NaN	50	NaN	NaN	NaN	2.7	NaN	67656.0	NaN	NaN	0.000	0.0

2457	Sudan	2015	Developing	64.1	225.0	58	0.0	0.000000	93.0	3585	NaN	85	93.0	NaN	93.0	0.3	2513.884661	3864783.0	NaN	NaN	0.488	7.2
2458	Sudan	2014	Developing	63.8	229.0	59	0.01	253.608651	94.0	676	NaN	86	94.0	8.43	94.0	0.3	2176.898290	37737913.0	NaN	NaN	0.485	7.2
2459	Sudan	2013	Developing	63.5	232.0	60	0.01	227.835321	93.0	2813	NaN	88	93.0	8.42	93.0	0.3	1955.667990	36649918.0	NaN	NaN	0.478	7.0
2460	Sudan	2012	Developing	63.2	235.0	61	0.01	220.522192	92.0	8523	NaN	89	92.0	8.20	92.0	0.3	1892.894352	3599192.0	NaN	NaN	0.468	6.8
2461	Sudan	2011	Developing	62.7	241.0	61	2.12	196.689215	93.0	5616	NaN	91	93.0	8.30	93.0	0.3	1666.857757	35167314.0	NaN	NaN	0.463	7.0
2462	Sudan	2010	Developing	62.5	243.0	62	1.77	172.009788	75.0	680	NaN	92	9.0	7.97	9.0	0.3	1476.478870	34385963.0	NaN	NaN	0.461	7.0
2463	Sudan	2009	Developing	62.0	248.0	63	1.99	17.053693	72.0	68	NaN	94	81.0	8.40	81.0	0.3	1226.884381	3365619.0	NaN	NaN	0.456	6.8
2464	Sudan	2008	Developing	61.8	251.0	64	2.01	128.636271	78.0	129	NaN	95	85.0	8.17	86.0	0.3	1291.528826	32955496.0	NaN	NaN	0.444	6.3
2465	Sudan	2007	Developing	61.4	254.0	65	2.01	86.131669	78.0	327	NaN	97	84.0	4.72	84.0	0.3	1115.695200	32282526.0	NaN	NaN	0.440	6.4
2466	Sudan	2006	Developing	61.0	260.0	66	1.9	60.336857	6.0	228	NaN	99	77.0	3.93	78.0	0.2	893.879364	316764.0	NaN	NaN	0.430	6.2
2467	Sudan	2005	Developing	67.0	261.0	66	1.55	37.590396	22.0	1374	NaN	101	78.0	3.18	78.0	0.2	679.753995	3911914.0	NaN	NaN	0.423	6.1
2468	Sudan	2004	Developing	59.7	278.0	68	1.59	37.044800	0.0	9562	NaN	102	74.0	3.39	74.0	0.2	565.569459	3186341.0	NaN	NaN	0.415	5.7
2469	Sudan	2003	Developing	59.6	278.0	69	1.74	35.352647	0.0	4381	NaN	104	69.0	3.18	69.0	0.2	477.738478	29435944.0	NaN	NaN	0.409	5.6
2470	Sudan	2002	Developing	59.4	277.0	70	1.59	30.622875	0.0	4529	NaN	106	6.0	2.95	6.0	0.2	412.151756	28679565.0	NaN	NaN	0.403	5.6
2471	Sudan	2001	Developing	58.9	283.0	71	1.81	28.880697	0.0	4362	NaN	108	66.0	2.96	66.0	0.2	377.525445	279455.0	NaN	NaN	0.399	5.6
2472	Sudan	2000	Developing	58.6	284.0	71	1.76	30.860010	0.0	2875	NaN	109	62.0	3.23	62.0	0.1	361.358430	2725535.0	NaN	NaN	0.394	5.5

ต่อไปเป็นการวิเคราะห์ข้อมูลแถวที่มีค่า BMI เป็น NULL โดยที่เราได้ว่า column BMI นั้น เป็นค่าดัชนีมวलय โดยที่ดัชนีมวलयนั้น จะไม่มีทางเป็นค่า 0 และไม่สามารถคำนวณจาก column อื่นได้ เราเลยเลือกที่จะลบแถวนั้นออกไป

```
new=data[data[' BMI '].isnull()]
data.drop(new.index,inplace=True)
```

โค้ดนี้จะเป็นการลบแถวที่ BMI นั้นเป็น NULL ทิ้งไป โดยที่จะสร้าง dataframe ใหม่มาอันนี้

โดยเลือกแถวที่มีคอลัมน์ BMI เป็น NULL จากนั้นจะ drop แถวใน data ที่มี index เดียวกันกับ new ออกไปโดยไม่ต้องกำหนด DataFrame ใหม่

```
data.isnull().any()
```

ต่อไปก็ดู column ที่เป็น NULL อีกรอบ

```

Country          False
Year             False
Status           False
Life expectancy  False
Adult Mortality  False
infant deaths     False
Alcohol          False
percentage expenditure  False
Hepatitis B      False
Measles          False
BMI              False
under-five deaths False
Polio            True
Total expenditure True
Diphtheria       True
HIV/AIDS         False
GDP              True
Population       True
thinness 1-19 years False
thinness 5-9 years False
Income composition of resources True
Schooling        True
dtype: bool

```

ข้อมูลที่ได้จาก `data.isnull().any()`

```

data[data['Polio'].isnull()]
data['Polio'].fillna('0.0', inplace=True)

```

ต่อไปเป็นการวิเคราะห์ข้อมูลที่แถวที่มีค่า Polio เป็น NULL โดยที่เราได้ดูว่า column Polio นั้นใช้อธิบายหน่วยการการฉีดวัคซีนป้องกัน Polio ในเด็กอายุ 1 ปี คิดเป็นเปอร์เซ็นต์ ซึ่งในตารางนี้ ไม่มีค่าที่เป็น 0.0 ซึ่งมันสามารถมีการฉีดวัคซีนป้องกันไวรัสตับอักเสบบีเป็น 0 เปอร์เซ็นต์ได้ ดังนั้นเราจึงเลือกวิธีการเปลี่ยนเป็น 0 ในช่องที่เป็น NULL

```

data.isnull().any()

```

ต่อไปก็ดู column ที่เป็น NULL อีกรอบ

```

Country          False
Year             False
Status           False
Life expectancy  False
Adult Mortality  False
infant deaths    False
Alcohol          False
percentage expenditure  False
Hepatitis B      False
Measles          False
BMI              False
under-five deaths      False
Polio            False
Total expenditure    True
Diphtheria        True
HIV/AIDS          False
GDP               True
Population        True
thinness 1-19 years  False
thinness 5-9 years  False
Income composition of resources  True
Schooling         True
dtype: bool

```

ข้อมูลที่ได้จาก data.isnull().any()

```

data[data['Total expenditure'].isnull()]

```

บรรทัดนี้ใช้ฟังก์ชัน isnull() เพื่อกรองแถวที่มีค่าว่างในคอลัมน์ Total expenditure

ผลลัพธ์ของโค้ดนี้จะเป็นชุดข้อมูลใหม่ที่ประกอบด้วยเฉพาะแถวที่มีค่าสำหรับ Total expenditure

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
32	Algeria	2015	Developing	75.6	19.0	21	0.0	0.0	95.0	63	59.5	24	95.0	NaN	95.0	0.1	4132.762920	39871528.0	6.0	5.8	0.743	14.4
48	Angola	2015	Developing	52.4	335.0	66	0.0	0.0	64.0	118	23.3	98	7.0	NaN	64.0	1.9	3695.793748	2785935.0	8.3	8.2	0.531	11.4
64	Antigua and Barbuda	2015	Developing	76.4	13.0	0	0.0	0.0	99.0	0	47.7	0	86.0	NaN	99.0	0.2	13566.954100	NaN	3.3	3.3	0.784	13.9
80	Argentina	2015	Developing	76.3	116.0	8	0.0	0.0	94.0	0	62.8	9	93.0	NaN	94.0	0.1	13467.123600	43417765.0	1.0	0.9	0.826	17.3
96	Armenia	2015	Developing	74.8	118.0	1	0.0	0.0	94.0	33	54.9	1	96.0	NaN	94.0	0.1	369.654776	291695.0	2.1	2.2	0.741	12.7
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2858	Venezuela (Bolivarian Republic of)	2015	Developing	74.1	157.0	9	0.0	0.0	87.0	0	62.1	10	87.0	NaN	87.0	0.1	NaN	NaN	1.6	1.5	0.769	14.3
2874	Viet Nam	2015	Developing	76.0	127.0	28	0.0	0.0	97.0	256	17.5	35	97.0	NaN	97.0	0.1	NaN	NaN	14.2	14.5	0.678	12.6
2890	Yemen	2015	Developing	65.7	224.0	37	0.0	0.0	69.0	468	41.3	47	63.0	NaN	68.0	0.1	NaN	NaN	13.6	13.4	0.499	9.0
2906	Zambia	2015	Developing	61.8	33.0	27	0.0	0.0	9.0	9	23.4	40	9.0	NaN	9.0	4.1	1313.889646	161587.0	6.3	6.1	0.576	12.5
2922	Zimbabwe	2015	Developing	67.0	336.0	22	0.0	0.0	87.0	0	31.8	32	88.0	NaN	87.0	6.2	118.693830	15777451.0	5.6	5.5	0.507	10.3

212 rows x 22 columns

ต่อไปเป็นการวิเคราะห์ข้อมูลแถวที่มีค่า Total expenditure เป็น NULL โดยที่เราได้ดูว่า column Total expenditure นั้นใช้อธิบายรายจ่ายด้านสุขภาพของรัฐบาลทั่วไปคิดเป็นเปอร์เซ็นต์ซึ่งไม่มีทางที่จะเป็นค่า NULL ได้อย่างแน่นอนเพราะอย่างน้อยรัฐบาลต้องมั่งมีงบประมาณเกี่ยวกับด้านนี้แน่นอน ดังนั้นเราจึงตัดสินใจที่จะลบแถวที่มีค่า Total expenditure เป็น NULL ออกไป

```
data[data['Total expenditure'].isnull()]
data.dropna(subset=['Total expenditure'], inplace=True)
```

```
data[data['Total expenditure'].isnull()]
```

โค้ดนี้ใช้ data เป็นตัวแปรที่อ้างอิงถึง Pandas DataFrame และ data['Total expenditure'] เป็นตัวแปรที่อ้างอิงถึงคอลัมน์ "Total expenditure" ใน DataFrame โค้ดนี้จะค้นหาแถวทั้งหมดใน DataFrame ที่ค่าในคอลัมน์ "Total expenditure" เป็นค่าว่าง ผลลัพธ์จะแสดงเป็น DataFrame ใหม่ที่ประกอบด้วยแถวเหล่านั้น

```
data.dropna(subset=['Total expenditure'], inplace=True)
```

โค้ดนี้ใช้ data.dropna() เมธอดเพื่อลบแถวที่มีค่า "Total expenditure" เป็นค่าว่างออกจาก DataFrame data โค้ดนี้ใช้พารามิเตอร์ subset=['Total expenditure'] เพื่อระบุว่าการลบแถวที่มีค่าว่างในคอลัมน์ "Total expenditure" เท่านั้น พารามิเตอร์ inplace=True บอกให้ dropna() เมธอดแก้ไข DataFrame data ดัชนีแบบแทนที่จะสร้าง DataFrame ใหม่

ที่เอาออกเพราะว่า เราไม่สามารถทราบได้ว่ารัฐบาลจะให้เงินค่ารักษาเท่าไร

```
data.isnull().any()
```

ต่อไปก็ดู column ที่เป็น NULL อีกรอบ

```
Country                False
Year                  False
Status                 False
Life expectancy        False
Adult Mortality        False
infant deaths          False
Alcohol                False
percentage expenditure False
Hepatitis B            False
Measles                False
BMI                   False
under-five deaths      False
Polio                  False
Total expenditure      False
Diphtheria             True
HIV/AIDS              False
GDP                    True
Population             True
thinness 1-19 years    False
thinness 5-9 years     False
Income composition of resources True
Schooling              True
dtype: bool
```

ข้อมูลที่ได้จาก data.isnull().any()

```
data[data['Diphtheria '].isnull()]
```

ต่อไปจะดูตารางในแถวที่มีค่า Diphtheria เป็น NULL ว่ามันเกิดจากอะไร

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
1742	Montenegro	2005	Developing	73.6	133.0	0	0.0	527.307672	0.0	0	55.7	0	0.0	8.46	NaN	0.1	3674.617924	614261.0	2.3	2.3	0.746	12.8
1743	Montenegro	2004	Developing	73.5	134.0	0	0.01	57.121901	0.0	0	55.0	0	0.0	8.45	NaN	0.1	338.199535	613353.0	2.3	2.4	0.740	12.6
1744	Montenegro	2003	Developing	73.5	134.0	0	0.01	495.078296	0.0	0	54.2	0	0.0	8.91	NaN	0.1	2789.173500	612267.0	2.4	2.4	0.000	0.0
1745	Montenegro	2002	Developing	73.4	136.0	0	0.01	36.480240	0.0	0	53.5	0	0.0	8.33	NaN	0.1	216.243274	69828.0	2.5	2.5	0.000	0.0
1746	Montenegro	2001	Developing	73.3	136.0	0	0.01	33.669814	0.0	0	52.7	0	0.0	8.23	NaN	0.1	199.583957	67389.0	2.5	2.6	0.000	0.0
1747	Montenegro	2000	Developing	73.0	144.0	0	0.01	274.547260	0.0	0	51.9	0	0.0	7.32	NaN	0.1	1627.428930	6495.0	2.6	2.7	0.000	0.0
2615	Timor-Leste	2001	Developing	59.4	269.0	3	0.5	6.556583	0.0	0	12.3	4	0.0	3.75	NaN	0.1	56.424987	892531.0	12.1	12.2	0.470	9.8
2616	Timor-Leste	2000	Developing	58.7	276.0	3	0.5	49.069672	0.0	0	11.9	4	0.0	3.26	NaN	0.1	422.286330	87167.0	12.2	12.2	0.000	0.0

ต่อไปเป็นการวิเคราะห์ข้อมูลที่แถวที่มีค่า Diphtheria เป็น NULL โดยที่เราได้ดูว่า column Diphtheria นั้นใช้อธิบายหน่วยการการฉีดวัคซีนป้องกันโรคคอตีบ บาดทะยัก และไอกรนในเด็กอายุ 1 ปีคิดเป็นเปอร์เซ็นต์ ซึ่งในตารางนี้ไม่มีค่าที่เป็น 0.0 ซึ่งมันสามารถมีการฉีดวัคซีนป้องกันวัคซีนป้องกันโรคคอตีบ บาดทะยัก และไอกรนเป็น 0 เปอร์เซ็นต์ได้ ดังนั้นเราจึงเลือกวิธีการเปลี่ยนเป็น 0 ในช่องที่เป็น NULL

```
data['Diphtheria '].fillna('0.0',inplace=True)
```

โค้ดนี้เป็นการเติมค่า NULL ในคอลัมน์ Diphtheria ด้วยค่า 0.0

```
data[data['GDP'].isnull()]
```

ต่อไปจะดูตารางในแถวที่มีค่า GDP เป็น NULL ว่ามันเกิดจากอะไร

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
161	Bahamas	2014	Developing	75.4	16.0	0	9.45	0.0	96.0	0	63.8	0	96.0	7.74	96.0	0.1	NaN	NaN	2.5	2.5	0.789	12.6
162	Bahamas	2013	Developing	74.8	172.0	0	9.42	0.0	97.0	0	63.2	0	97.0	7.50	97.0	0.1	NaN	NaN	2.5	2.5	0.790	12.6
163	Bahamas	2012	Developing	74.9	167.0	0	9.5	0.0	96.0	0	62.6	0	99.0	7.43	98.0	0.2	NaN	NaN	2.5	2.5	0.789	12.6
164	Bahamas	2011	Developing	75.0	162.0	0	9.34	0.0	95.0	0	62.0	0	97.0	7.63	98.0	0.1	NaN	NaN	2.5	2.5	0.788	12.6
165	Bahamas	2010	Developing	75.0	161.0	0	9.19	0.0	98.0	0	61.3	0	97.0	7.44	99.0	0.2	NaN	NaN	2.5	2.5	0.788	12.6
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2901	Yemen	2004	Developing	62.2	247.0	42	0.06	0.0	43.0	12708	33.3	56	72.0	4.90	72.0	0.1	NaN	NaN	13.9	13.9	0.464	8.4
2902	Yemen	2003	Developing	61.9	249.0	43	0.04	0.0	38.0	8536	32.7	58	61.0	5.00	61.0	0.1	NaN	NaN	14.0	13.9	0.457	8.2
2903	Yemen	2002	Developing	61.5	25.0	45	0.07	0.0	31.0	890	32.2	61	64.0	4.22	65.0	0.1	NaN	NaN	14.0	14.0	0.450	8.0
2904	Yemen	2001	Developing	61.1	251.0	46	0.08	0.0	19.0	485	31.7	63	73.0	4.34	73.0	0.1	NaN	NaN	14.0	14.0	0.444	7.9
2905	Yemen	2000	Developing	68.0	252.0	48	0.07	0.0	14.0	0	31.2	66	74.0	4.14	74.0	0.1	NaN	NaN	14.1	14.1	0.436	7.7

375 rows x 22 columns

ต่อไปเป็นการวิเคราะห์ข้อมูลที่แถวที่มีค่า GDP เป็น NULL โดยที่เราได้ดูว่า column GDP นั้นใช้อธิบายผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวซึ่งไม่มีทางที่จะเป็นค่า NULL ได้อย่างแน่นอน ดังนั้นเราจึงตัดสินใจที่จะลบแถวที่มีค่า GDP เป็น NULL ออกไป

```
new=data[data['GDP'].isnull()]\nnew.drop(new.index,inplace=True)
```

new = data[data['GDP'].isnull()]

บรรทัดนี้สร้าง DataFrame ใหม่ชื่อ new ซึ่งประกอบด้วยแถวทั้งหมดจาก DataFrame data ที่มีค่า GDP เป็น null

```
data.drop(new.index, inplace=True)
```

บรรทัดนี้ลบแถวทั้งหมดจาก DataFrame data ที่มีดัชนีอยู่ใน DataFrame new

```
data.isnull().any()
```

ต่อไปก็ดู column ที่เป็น NULL อีกรอบ

```
Country      False
Year          False
Status        False
Life expectancy  False
Adult Mortality False
infant deaths  False
Alcohol        False
percentage expenditure False
Hepatitis B    False
Measles        False
BMI            False
under-five deaths False
Polio          False
Total expenditure False
Diphtheria     False
HIV/AIDS      False
GDP            False
Population    True
thinness 1-19 years False
thinness 5-9 years False
Income composition of resources False
Schooling      False
dtype: bool
```

ข้อมูลที่ได้จาก data.isnull().any()

```
data.dropna(subset=['Population'], inplace=True)
```

เราไม่สามารถวัดจำนวนประชากรจากข้อมูลที่เรามีอยู่ได้

```
data.isnull().any()
```

ต่อไปก็ดู column ที่เป็น NULL อีกรอบ

```
data.to_csv("data.csv")
```


นำ ข้อมูลจาก data ไปทำเป็นไฟล์ csv

```
data.info()
```

ต่อไปจะดูข้อมูลทางเทคนิคของ dataframe เช่น จำนวนแถวและคอลัมน์ทั้งหมด ชนิดของข้อมูลในแต่ละคอลัมน์

```
<class 'pandas.core.frame.DataFrame'>
Index: 2102 entries, 0 to 2937
Data columns (total 22 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Country                               2102 non-null   object
1   Year                                  2102 non-null   int64
2   Status                                2102 non-null   object
3   Life expectancy                       2102 non-null   float64
4   Adult Mortality                       2102 non-null   float64
5   infant deaths                         2102 non-null   int64
6   Alcohol                               2102 non-null   object
7   percentage expenditure                2102 non-null   float64
8   Hepatitis B                           2102 non-null   object
9   Measles                               2102 non-null   int64
10  BMI                                    2102 non-null   float64
11  under-five deaths                     2102 non-null   int64
12  Polio                                 2102 non-null   object
13  Total expenditure                     2102 non-null   float64
14  Diphtheria                            2102 non-null   object
15  HIV/AIDS                              2102 non-null   float64
16  GDP                                    2102 non-null   float64
17  Population                             2102 non-null   float64
18  thinness 1-19 years                   2102 non-null   float64
19  thinness 5-9 years                    2102 non-null   float64
20  Income composition of resources        2102 non-null   float64
21  Schooling                             2102 non-null   float64
dtypes: float64(12), int64(4), object(6)
memory usage: 377.7+ KB
```





```
def clean_column_name(column_name):  
    column_name = column_name.strip()  
    column_name = column_name.replace(' ', '_')  
    column_name = column_name.title()  
    return column_name  
  
df = df.rename(columns=clean_column_name)
```

ทำการแก้ ชื่อ columns เพื่อให้ทำงานง่ายต่อการ Train Model

## EDA and visualization of data

```
life.info()

<class 'pandas.core.frame.DataFrame'>
Index: 2102 entries, 0 to 2937
Data columns (total 22 columns):
#   Column                                Non-Null Count  Dtype  
---  -
0   Country                                2102 non-null   object  
1   Year                                  2102 non-null   int64   
2   Status                                2102 non-null   object  
3   Life_Expectancy                       2102 non-null   float64  
4   Adult_Mortality                       2102 non-null   float64  
5   Infant_Deaths                         2102 non-null   int64   
6   Alcohol                               2102 non-null   float64  
7   Percentage_Expenditure                2102 non-null   float64  
8   Hepatitis_B                           2102 non-null   float64  
9   Measles                               2102 non-null   int64   
10  Bmi                                    2102 non-null   float64  
11  Under-Five_Deaths                     2102 non-null   int64   
12  Polio                                 2102 non-null   float64  
13  Total_Expenditure                     2102 non-null   float64  
14  Diphtheria                            2102 non-null   float64  
15  Hiv/Aids                              2102 non-null   float64  
16  Gdp                                    2102 non-null   float64  
17  Population                            2102 non-null   float64  
18  Thinness_1-19_Years                   2102 non-null   float64  
19  Thinness_5-9_Years                    2102 non-null   float64  
20  Income_Composition_Of_Resources        2102 non-null   float64  
21  Schooling                             2102 non-null   float64  
dtypes: float64(16), int64(4), object(2)
memory usage: 377.7+ KB

[ ] life.drop('Country', axis=1, inplace=True)
    life.drop('Year', axis=1, inplace=True)
    life.drop('Status', axis=1, inplace=True)
```

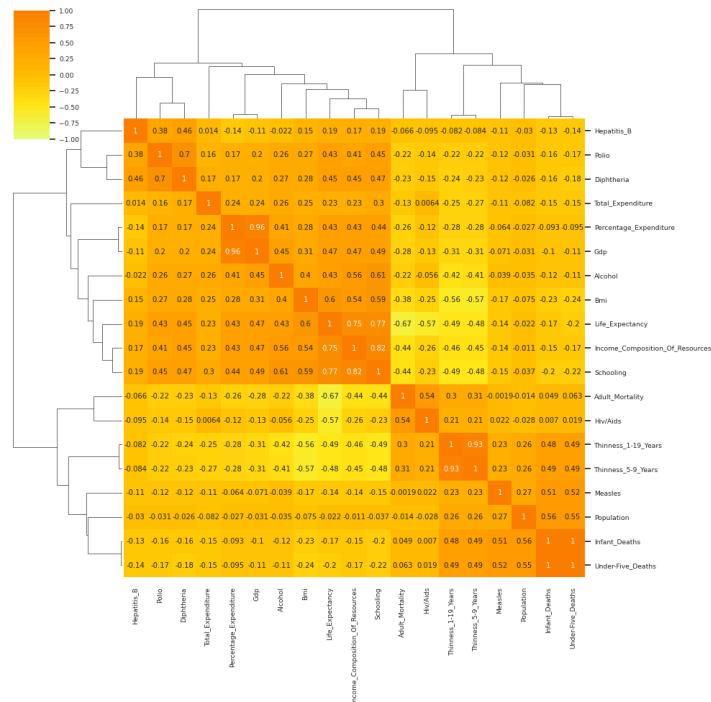
หลังจากเราได้ทำการ cleaning data แล้ว จะทำการ drop columns ที่มี data type เป็น object ให้ data set มีเพียงข้อมูล ที่เป็น int และ float เพื่อนำไปหา correlation ในขั้นนี้เรา drop column Country , Year และ Status

```
{ } life.corr()

Life_Expectancy  Adult_Mortality  Infant_Deaths  Alcohol  Percentage_Expenditure  Hepatitis_B  Measles  Bmi  Under-Five_Deaths  Polio  Total_Expenditure  Diphtheria  Hiv/Aids  Gdp  Population
Life_Expectancy    1.000000    -0.672928    -0.171120    0.429032    0.428922    0.189429    -0.140390    0.597377    -0.196737    0.426739    0.233668    0.449848    -0.572709    0.465955    -0.022220
Adult_Mortality    -0.672928    1.000000    0.048754    -0.216136    -0.262638    -0.065946    -0.001904    -0.381666    0.063221    -0.224261    -0.125904    -0.225058    0.539379    -0.283133    -0.014397
Infant_Deaths      -0.171120    0.048754    1.000000    -0.116235    -0.092846    -0.128399    0.508678    -0.230923    0.996742    -0.156345    -0.151100    -0.159548    0.007044    -0.103120    0.563791
Alcohol            0.429032    -0.216136    -0.116235    1.000000    0.412093    -0.021812    -0.038761    0.398463    -0.112177    0.262617    0.255722    0.267179    -0.056067    0.453114    -0.035054
Percentage_Expenditure  0.428922    -0.262638    -0.092846    0.412093    1.000000    -0.135855    -0.064145    0.275338    -0.095461    0.168842    0.238926    0.169564    -0.117443    0.957672    -0.027161
Hepatitis_B        0.189429    -0.065946    -0.128399    -0.021812    -0.135855    1.000000    -0.108079    0.145707    -0.140927    0.375538    0.014207    0.457136    -0.095201    -0.112996    -0.030357
Measles            -0.140390    -0.001904    0.508678    -0.038761    -0.064145    -0.108079    1.000000    -0.173981    0.518045    -0.115027    -0.111896    -0.120254    0.022319    -0.071079    0.269704
Bmi                0.597377    -0.381666    -0.230923    0.398463    0.275338    0.145707    -0.173981    1.000000    -0.241206    0.274313    0.247288    0.277622    -0.245260    0.306628    -0.075230
Under-Five_Deaths  -0.196737    0.063221    0.996742    -0.112177    -0.095461    -0.140927    0.518045    -0.241206    1.000000    -0.173974    -0.151729    -0.180035    0.018726    -0.106635    0.549806
Polio              0.426739    -0.224261    -0.156345    0.262617    0.168842    0.375538    -0.115027    0.274313    -0.173974    1.000000    0.163360    0.695677    -0.140017    0.197313    -0.031031
Total_Expenditure  0.233668    -0.125904    -0.151100    0.255722    0.238926    0.014207    -0.111896    0.247288    -0.151729    0.163360    1.000000    0.174435    0.006447    0.236585    -0.081765
Diphtheria         0.449848    -0.225058    -0.159548    0.267179    0.169564    0.457136    -0.120254    0.277622    -0.180035    0.695677    0.174435    1.000000    -0.149118    0.195240    -0.025871
Hiv/Aids           -0.572709    0.539379    0.007044    -0.056067    -0.117443    -0.095201    0.022319    -0.245260    0.018726    -0.140017    0.006447    -0.149118    1.000000    -0.131439    -0.028486
Gdp                0.465955    -0.283133    -0.103120    0.453114    0.957672    -0.112996    -0.071079    0.306628    -0.106635    0.197313    0.236585    0.195240    -0.131439    1.000000    -0.030587
Population         -0.022220    -0.014397    0.563791    -0.035054    -0.027161    -0.030357    0.269704    -0.075230    0.549806    -0.031031    -0.081765    -0.025871    -0.028486    -0.030587    1.000000
Thinness_1-19_Years -0.485611    0.303865    0.483636    -0.423803    -0.282519    -0.082155    0.229773    -0.564588    0.486185    -0.222943    -0.254290    -0.240375    0.205399    -0.308819    0.262685
Thinness_5-9_Years -0.478483    0.310002    0.489351    -0.413761    -0.284505    -0.084091    0.225011    -0.571038    0.490715    -0.224492    -0.268919    -0.231912    0.208411    -0.310562    0.260300
Income_Composition_Of_Resources  0.749653    -0.440299    -0.149267    0.562073    0.426417    0.173336    -0.141301    0.544893    -0.168532    0.412032    0.234323    0.447336    -0.256278    0.473724    -0.010933
Schooling          0.765871    -0.435537    -0.204528    0.613631    0.440187    0.185019    -0.151279    0.587406    -0.220797    0.448793    0.297371    0.465387    -0.225909    0.488736    -0.036942
```

ภาพนี้แสดงถึงค่า correlation ของทุก attributed

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(font_scale=0.6)
sns.clustermap(life.corr(),annot=True ,cmap='Wistia', vmin=-1, vmax=1)
plt.show()
```

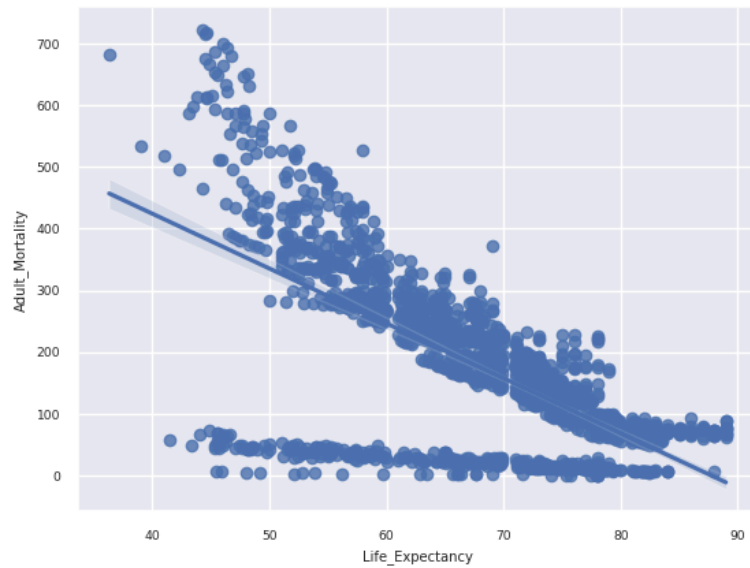


ทำการใช้ Seaborn และ Matplotlib เพื่อทำ heat map มาดูค่า correlation ที่เกี่ยวข้องกับสิ่งที่เราต้องการหา ในที่นี้ เราต้องการหา correlation ที่สัมพันธ์กับ Life\_Expectancy ทุกตัวที่มีค่า correlation มากกว่า 0.5 แต่ไม่ถึง 1 และ น้อยกว่า -0.5 แต่ไม่ถึง -1 เพื่อให้ได้เป็น positive relationship และ negative relationship

จาก correlation heatmap ทำให้เราทราบว่า มี attributed Adult\_Mortality, Bmi, Hiv/Aids, Income\_Composition\_Of\_Resources และ Schooling ที่มี correlation ที่สัมพันธ์กับ Life\_Expectancy และมีค่าตรงกับที่เราต้องการ จากนั้นนำ attributed ที่เราได้มาไปทำ scatter plot กับ Life\_Expectancy



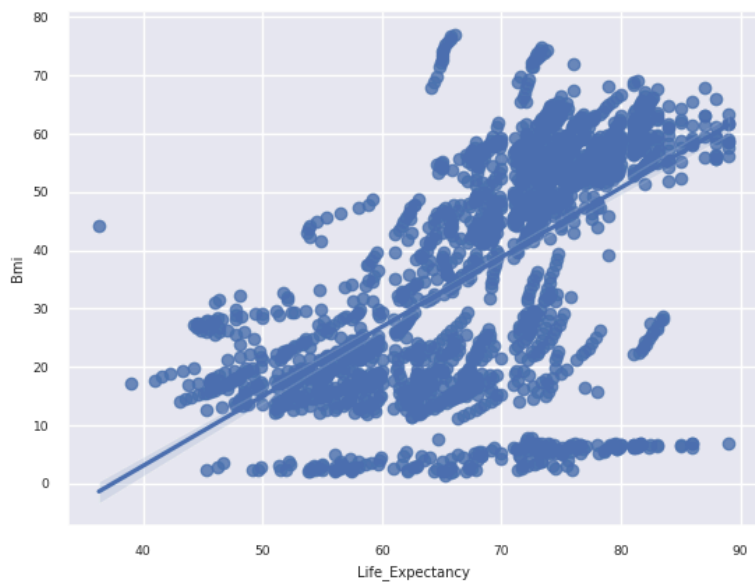
```
c = life['Life_Expectancy'].corr(life['Adult_Mortality'])
print("Correlation = %0.2f"%life['Life_Expectancy'].corr(life['Adult_Mortality']))
sns.regplot(data=life, x='Life_Expectancy', y='Adult_Mortality')
plt.show()
```



Scatter plot Negative relationship , Adult\_Mortality vs Life\_Expectancy



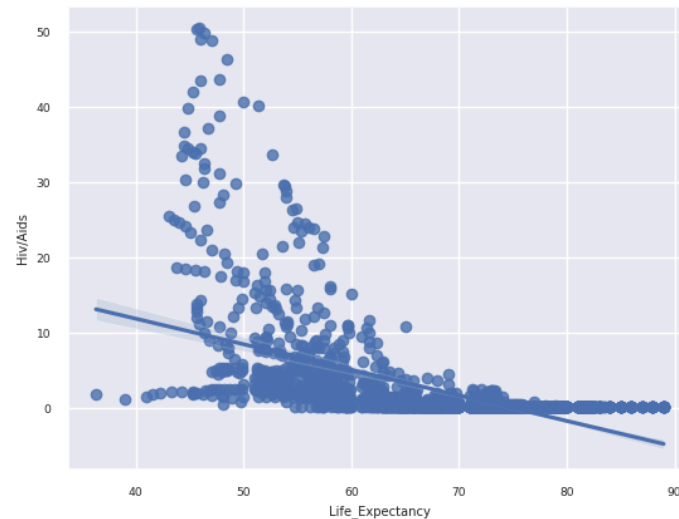
```
c = life['Life_Expectancy'].corr(life['Bmi'])
print("Correlation = %0.2f"%life['Life_Expectancy'].corr(life['Bmi']))
sns.regplot(data=life, x='Life_Expectancy', y='Bmi')
plt.show()
```



Scatter plot Positive relationship , Bmi vs Life\_Expectancy



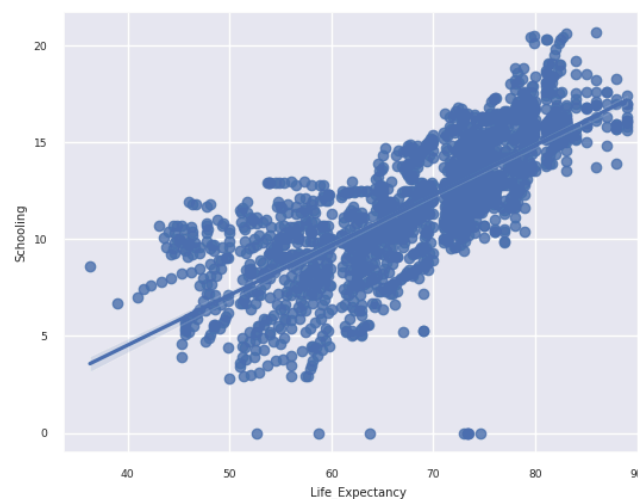
```
c = life['Life_Expectancy'].corr(life['Hiv/Aids'])
print("Correlation = %0.2f"%life['Life_Expectancy'].corr(life['Hiv/Aids']))
sns.regplot(data=life, x='Life_Expectancy', y='Hiv/Aids')
plt.show()
```



Scatter plot Negative relationship , Hiv/Aids vs Life\_Expectancy



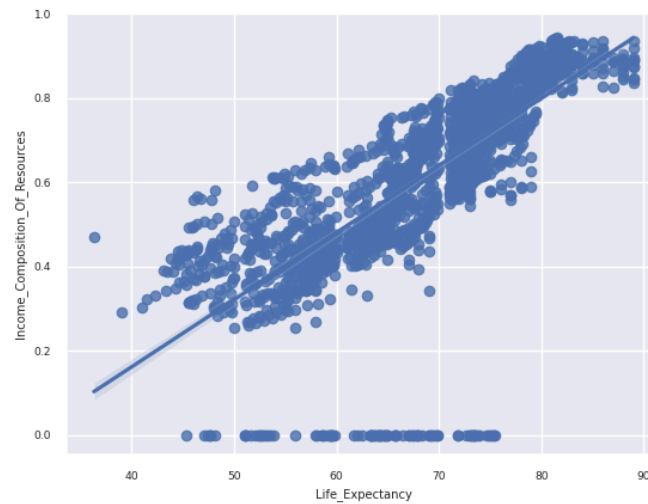
```
c = life['Life_Expectancy'].corr(life['Schooling'])
print("Correlation = %0.2f"%life['Life_Expectancy'].corr(life['Schooling']))
sns.regplot(data=life, x='Life_Expectancy', y='Schooling')
plt.show()
```



Scatter plot Positive relationship , Schooling vs Life\_Expectancy



```
c = life['Life_Expectancy'].corr(life['Income_Composition_Of_Resources'])
print("Correlation = %0.2f"%life['Life_Expectancy'].corr(life['Income_Composition_Of_Resources']))
sns.regplot(data=life, x='Life_Expectancy', y='Income_Composition_Of_Resources')
plt.show()
```



Scatter plot Positive relationship , Income\_Composition\_Of\_Resources vs Life\_Expectancy

จากการพิจารณา Scatter plot ที่เรามาจะนำเอา attributed ที่สัมพันธ์กับ Life\_Expectancy ทั้งหมดเพื่อไปทำขั้นตอนการ Modeling โดยเลือกใช้เป็น Linear regression

## Modeling method

เราสามารถใช้ linear regression เพื่อที่จะมาทำนาย life expectancy(อายุขัย)

โดยเริ่มต้นที่ import library คือ

1.numpy

2.ols (ordinary least squares regression)

3.mean\_absolute\_error, mean\_squared\_error, r2\_score

```
import numpy as np
from statsmodels.formula.api import ols
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

ถัดมาทำการสร้าง Model 1 โดยเลือกใช้ Attribute ที่มีค่า Correlation มากกว่า 50 เมื่อนำมาเทียบกับ Attribute Life Expectancy และ ทำการแสดงผลโมเดล

```
life.rename(columns={'Hiv/Aids': 'HivAids'}, inplace=True)

model1 = ols('Life_Expectancy ~ Adult_Mortality+Bmi+HivAids+Schooling+Income_Composition_Of_Resources', life).fit()
model1.summary()
```

OLS Regression Results

Dep. Variable:	Life_Expectancy	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.820
Method:	Least Squares	F-statistic:	1916.
Date:	Mon, 13 May 2024	Prob (F-statistic):	0.00
Time:	06:49:19	Log-Likelihood:	-5992.9
No. Observations:	2102	AIC:	1.200e+04
Df Residuals:	2096	BIC:	1.203e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	51.4395	0.452	113.893	0.000	50.554	52.325
Adult_Mortality	-0.0168	0.001	-18.423	0.000	-0.019	-0.015
Bmi	0.0560	0.006	9.564	0.000	0.045	0.068
HivAids	-0.4859	0.019	-26.115	0.000	-0.522	-0.449
Schooling	1.0390	0.051	20.348	0.000	0.939	1.139
Income_Composition_Of_Resources	10.7542	0.783	13.731	0.000	9.218	12.290

Omnibus:	130.089	Durbin-Watson:	0.616
Prob(Omnibus):	0.000	Jarque-Bera (JB):	565.576
Skew:	-0.020	Prob(JB):	1.54e-123
Kurtosis:	5.541	Cond. No.	1.85e+03

ถัดมาหลังจากที่เราได้ Model ที่ 1 เราทำการเช็คค่า

- 1.R Squared
- 2.Mean Absolute Error
- 3.Mean Squared Error
- 4.root Mean Squared Error

เพื่อตรวจสอบความแม่นยำของ Model นี้

```
# Predicting the target variable 'Life_expectancy' using the model
predicted_life_expectancy = model1.predict(life[['Adult_Mortality','Bmi','HivAids','Schooling','Income_Composition_Of_Resources']])

# Calculating R-squared
r_squared = r2_score(life['Life_Expectancy'], predicted_life_expectancy)

# Calculating Mean Absolute Error (MAE)
mae = mean_absolute_error(life['Life_Expectancy'], predicted_life_expectancy)

# Calculating Mean Squared Error (MSE)
mse = mean_squared_error(life['Life_Expectancy'], predicted_life_expectancy)

# Calculating Root Mean Squared Error (RMSE)
rmse = np.sqrt(mse)

print("R-squared:", r_squared)
print("Mean Absolute Error:", mae)
print("Mean Squared Error:", mse)
print("Root Mean Squared Error:", rmse)
```

โดยที่ Model 1 ได้ผลลัพธ์ดังนี้

```
R-squared: 0.8205156260201248
Mean Absolute Error: 3.1120271603328495
Mean Squared Error: 17.53593317517557
Root Mean Squared Error: 4.187592766157613
```



ถัดมาทำการสร้าง Model 2 โดยเลือกใช้ Attribute ที่มีค่า Correlation มากกว่า 70 เมื่อนำมาเทียบกับ Attribute Life Expectancy และ ทำการแสดงผลโมเดล

```
model2 = ols('Life_Expectancy ~ Schooling+Income_Composition_Of_Resources', life).fit()  
model2.summary()
```

```
OLS Regression Results  
  
Dep. Variable: Life_Expectancy    R-squared: 0.630  
Model: OLS                      Adj. R-squared: 0.630  
Method: Least Squares          F-statistic: 1789.  
Date: Mon, 13 May 2024          Prob (F-statistic): 0.00  
Time: 06:57:34                  Log-Likelihood: -6752.4  
No. Observations: 2102          AIC: 1.351e+04  
Df Residuals: 2099              BIC: 1.353e+04  
Df Model: 2  
Covariance Type: nonrobust  
  
                coef  std err   t    P>|t| [0.025 0.975]  
Intercept      41.5937  0.488   85.266 0.000   40.637  42.550  
Schooling       1.3706  0.070   19.692 0.000    1.234   1.507  
Income_Composition_Of_Resources 17.3489  1.101   15.756 0.000   15.189  19.508  
Omnibus: 169.020 Durbin-Watson: 0.314  
Prob(Omnibus): 0.000 Jarque-Bera (JB): 664.951  
Skew: -0.306      Prob(JB): 4.05e-145  
Kurtosis: 5.687    Cond. No. 104.
```

ถัดมาหลังจากที่เราได้ Model ที่ 2 เราทำการเช็คค่า

1.R Squared

2.Mean Absolute Error

3.Mean Squared Error

4.root Mean Squared Error

เพื่อตรวจสอบความแม่นยำของ Model นี้โดยใช้โค้ดแบบเดียวกับ Model1

โดยที่ Model 2 ได้ผลลัพธ์ดังนี้

```
R-squared: 0.6302833167641191  
Mean Absolute Error: 4.381143497351642  
Mean Squared Error: 36.12195818059854  
Root Mean Squared Error: 6.010154588743832
```

## Modeling results and discussion

หลังจากการทดลองได้ผลลัพธ์ดังนี้

1. Model 1 มีผลลัพธ์ดังนี้  
R-squared: 0.8205156260201248  
Mean Absolute Error: 3.1120271603328495  
Mean Squared Error: 17.53593317517557  
Root Mean Squared Error: 4.187592766157613
2. Model 2 มีผลลัพธ์ดังนี้  
R-squared: 0.6302833167641191  
Mean Absolute Error: 4.381143497351642  
Mean Squared Error: 36.12195818059854  
Root Mean Squared Error: 6.010154588743832

จากผลลัพธ์การทดลองทำให้เห็นถึงความแตกต่างของ Model ทั้ง 2

## Conclusion

หลังจากทำการทดลองใช้ ols (ordinary least squares regression) เพื่อในการทำนาย

Life Expectancy(อายุขัย) จากข้อมูลที่ได้มาโดยทำการสร้าง Model 2 ตัวที่มีการเลือก Correlation ที่ต่างกันโดย Model 1 เลือกค่า Correlation ที่มากกว่า 0.5 และ Model 2 เลือกค่า Correlation ที่มากกว่า 0.7 โดยได้ผลลัพธ์คือ Model 1 ที่มีจำนวน Attribute มากกว่าเพราะจากช่วงของค่า Correlation ที่มากกว่า 0.5 ทำให้มี Attribute มากถึง 5 ตัว นั้นมีผลลัพธ์ที่ดีกว่าในทุก ๆ ด้านโดยเช็คจากค่า

1.R Squared

2.Mean Absolute Error

3.Mean Squared Error

4.root Mean Squared Error